

# Notes on Calculus-Based Probability Theory

ENGR 113 – Managerial Statistics

## 1 Background

The text by Berenson, Levine, and Krehbiel is a good introduction to the basic concepts of probability and statistics for students oriented toward careers in business and management. While it covers a wide range of topics and includes many good examples, the mathematical content is fairly low. For someone going into a career in a technical field, it is useful to have a deeper understanding of and skill in probability. Hence these notes will supplement the text with calculus-based probability.

Probability theory (and the accompanying statistical theory) is traditionally taught at one of three levels: no calculus, with calculus, or with measure theory. A traditional course for someone in a non-technical field would not use calculus. A first course for economics majors would normally contain some calculus, with more calculus in subsequent courses. An undergraduate course for majors in science, engineering, or math would be entirely calculus-based. An upper-level graduate course for students in math, statistics, or economics would be based on measure theory, a sort of “advanced calculus”. As information systems management is a blend of economics and computer science, with more emphasis on management preparation, this course will be somewhere in between all of those areas, based mostly on non-calculus math, but some knowledge of calculus-based probability is essential.

## 2 Review of Discrete Random Variables

Recall that a discrete random variable is one whose value will be from a finite set (e.g., how many times a coin shows heads when flipped 10 times) or a sequence of possible outcomes (e.g., the number of consecutive days you have to play a daily lottery until you win; note that there is no upper limit to how long this could take, but the possible values are all positive integers). Each of the possible values is accompanied by a probability, i.e.,  $P(X = x_i) = p_i$ . Since probabilities are always non-negative and must add to one, we have the following two conditions for probabilities:

1.  $p_i \geq 0$
2.  $\sum_i p_i = 1$ .

Discrete probability distributions are typically represented by either a table or a formula, or sometimes as a graph. For example, suppose you have two roommates, and in the morning, they often drink coffee. Suppose each one decides independently whether or not they will have coffee with probability .75. Then the number of your roommates who drink coffee on a particular morning is distributed as a binomial random variable with  $n = 2$  and  $p = .75$ . The possible values are  $\{0, 1, 2\}$  and the associated probabilities are the  $p_i$ , with  $p_0 = P(0 \text{ coffee drinkers}) = (.25) \cdot (.25) = .0625$ ,  $p_2 = P(\text{both roommates drink coffee}) = (.75) \cdot (.75) = .5625$ , and  $p_1 = P(\text{exactly one roommate drinks coffee}) = 1 - .5625 - .0625 = .3750$ .

As a table:

$x$	$P(X = x)$
0	.0625
1	.3750
2	.5625

As a formula:

$$P(X = x) = p_i = \begin{cases} \binom{2}{x} \cdot \left(\frac{3}{4}\right)^x \cdot \left(\frac{1}{4}\right)^{2-x} & \text{if } x = 0, 1, \text{ or } 2 \\ 0 & \text{otherwise.} \end{cases}$$

## 2.1 Summary Measures

Often a probability distribution is sufficiently complicated that it is easier to use a few summary measures to describe the distribution, rather than worrying about its full-blown detailed structure. As we saw earlier in this course, some of the important features to identify are the location, spread, and shape. Typically we use the mean (expected value) to describe location, and the standard deviation for spread. Shape is often best summarized by a graph.

Recall the definitions:

$$E[X] = \sum_{i=1}^{\infty} x_i p_i$$

$$Var(X) = E[(X - E[X])^2] = E[X^2] - (E[X])^2 = \sum_{i=1}^{\infty} x_i^2 p_i - \left( \sum_{i=1}^{\infty} x_i p_i \right)^2$$

$$\sigma_X = \sqrt{Var(X)}.$$

How do we interpret these summary measures? Here we give an explanation in terms of the relative frequency definition of probability (e.g., for an event that can happen over and over independently, the probability that this event happens is the fraction of time it actually does happen in the long run). Suppose our random variable  $X$  comes from an experiment that can be repeated independently many times. Say we repeat it  $N$  times, where  $N$  is large. Each time we do the experiment,  $X$  will take some value. By the relative frequency definition, if  $N$  is very large, then the fraction of times that  $X$  takes a particular value  $x_i$  will be close to  $p_i$ . Now if we take all  $N$  observations and average them,  $\bar{X} = \frac{1}{N}(X_1 + X_2 + \dots + X_N)$ . The  $N$  outcomes appear in a highly random order, but if we rearrange them, the average can be written as  $\sum_{i=1}^{\infty} x_i \cdot (\text{frequency of } x_i)/N = \sum_{i=1}^{\infty} x_i \cdot (\text{relative frequency of } x_i)$  which converges to  $\sum_{i=1}^{\infty} x_i p_i = E[X]$  as  $N \rightarrow \infty$ . Thus  $E[X]$  represents the long run average value of  $X$  if the experiment is repeated many times. Note that  $E[X]$  is not necessarily any one of the possible values,  $x_i$ , thus the name *expected value* can often be the cause of some confusion.

The variance is the expected squared distance of an observation from its expected value. Basically it is just a measure of the variability of the possible outcomes about their long run average. The standard deviation,  $\sigma_X$ , is the square root of the variance, and it is also a measure of spread, having the same units as  $X$  does. If the standard deviation is small, then most of the probability mass is very near to  $E[X]$ , whereas if the standard deviation is large, the mass is much more spread out. When the standard deviation is 0 (the smallest possible value), then the probability distribution is as concentrated as possible, namely all of the mass sits at a single location,  $E[X]$ , which means that  $X$  is a constant.

There are a few rules that are useful for making computations associated with means, variances and standard deviations. Let  $a$  and  $b$  be constants, with  $X$  being the random variable:

$$E[aX + b] = aE[X] + b$$

$$\sigma_{aX+b}^2 = a^2 \sigma_X^2$$

$$\sigma_{aX+b} = |a| \sigma_X$$

Adding any constant to a random variable will move all of its possible values by that constant amount, but the shape will remain unchanged. Consequently the spread will remain unchanged. Thus, the variance and standard deviation are not affected by adding or subtracting a constant to a random variable.

### 3 Continuous Random Variables

In many real-life situations, random quantities may be able to take any possible value on the real number line (or the positive part of it). For example, when waiting for something to happen (measuring time), it could take any positive amount of time to happen (although we may only be able to measure it to the nearest nanosecond, it is mathematically and practically convenient to think of it as real-valued rather than discrete, since there would be so many different possible values that it would be impossible to compute anything). When there are an infinite number of possible outcomes, it becomes much easier to describe the probability distribution using calculus techniques.

To understand how we can define a probability distribution for a continuous random variable, we must recognize that rather than having masses of probability associated with a discrete set of possible values, we want to have all the probability spread over a continuum of possible values. No lumps of probability should remain. But what happens to the probability? Is any or all of it lost? Indeed, none of it is lost. Rather, this spreading results in a *density of probability*, where the total probability of 1 is spread out over an interval of real numbers (or the whole real line). In particular, there will be a function,  $f(x)$ , called a *probability density function* or *pdf*, which defines the relative likelihood of each part of the interval. For discrete probability distributions, the probability of an event (such as  $X < 4$ ) is computed by adding up all of the probability mass described by the event (the set of possible outcomes). For continuous probability distributions, this probability will be computed by *integrating* the pdf, that is by finding the area under the pdf over the range of  $x$  values specified by the event. The total area under the pdf must account for all of the probability and hence must equal 1. For example, suppose we want to describe a probability distribution corresponding to the experiment “pick a random number between 0 and 1”. All real numbers in the interval  $[0,1]$  are possible outcomes, all real numbers outside  $[0,1]$  are impossible outcomes, and the numbers in  $[0,1]$  should be “equally likely”. To satisfy the “equally likely” requirement, we would spread the total probability of 1 uniformly over the interval of possible values,  $[0,1]$ . Thus the pdf would be given by:

$$f(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

Notice that all of the probability is concentrated on  $[0,1]$ , and that the area under the pdf  $f(x)$  is equal to 1. Probability density functions,  $f(x)$ , must satisfy two fundamental properties:

1.  $f(x) \geq 0$ ,  $-\infty < x < \infty$
2.  $\int_{-\infty}^{\infty} f(x)dx = 1$ .

These two properties associated with probability density functions are analogous to the two properties for discrete probability distributions, names  $p_i \geq 0$  and  $\sum_i p_i = 1$ .

Students are often initially confused by the interpretation of a probability density function,  $f(x)$ . Because we have spread the probability over an interval of possible values, *no single value has positive probability*, that is  $P(X = x) = 0$  for every  $x$ . While every single point in an interval

has zero probability, an *interval* of possible values can have positive probability. (Such is the magic of calculus!) Moreover, that probability is obtained by integrating the pdf over the interval of interest.

Specifically, suppose we have a variable  $X$ , and we want to compute  $P(a < X < b)$ . If  $X$  were discrete, then we would just sum up all of the  $p_i$  corresponding to any  $x_i$  satisfying  $a < x_i < b$ . In the continuous case, we do the continuous version of addition, namely integration, and  $P(a < X < b) = \int_a^b f(x)dx$ .

The probability density function specifies the *density* of probability at a point  $x$ , not the *amount* of probability at  $x$ . By density of probability, we mean the *average* amount of probability in a small region about  $x$ , or  $P(x \leq X \leq x + dx)/dx$ , the amount of probability in a small interval about  $x$  (the interval  $[x, x + dx]$ ) divided by the length of that interval. Because  $P(x \leq X \leq x + dx)/dx = (\int_x^{x+dx} f(u)du)/dx$ , the Fundamental Theorem of Calculus tells us that this density of probability must converge to  $f(x)$  as  $dx \rightarrow 0$ . Thus we can compute approximately  $P(x \leq X \leq x+h) \approx f(x) \cdot h$  for small values of  $h$ . The probability that  $X$  lies in a small interval including  $x$  is approximately the value of the pdf,  $f(x)$ , time the length of the interval,  $h$ . Of course,  $P(x \leq X \leq x+h)$  is computed exactly by  $\int_x^{x+h} f(u)du$ . Note that since  $P(X = a) = P(X = b) = 0$ ,  $P(a < X < b) = P(a \leq X \leq b)$  so we don't need to worry about whether a " $<$ " or a " $\leq$ " is used in describing events associated with a continuous random variable  $X$ , because the event probabilities will be the same.

Before giving some specific examples and discussing special distributions, let us mention how to compute expected values and variances for continuous random variables. Recall that for discrete random variables,  $E[X] = \sum_{i=1}^{\infty} x_i p_i$ , the sum of each possible value weighted by its probability. In the continuous case, the sum changes to an integral and we integrate over all possible values. However, instead of weighting by probabilities, we weight by the probability density function, the pdf:

$$E[X] = \int_{-\infty}^{\infty} x f(x) dx$$

$$Var(X) = E[X^2] - (E[X])^2 = \int_{-\infty}^{\infty} x^2 f(x) dx - \left( \int_{-\infty}^{\infty} x f(x) dx \right)^2.$$

### Example 1

Suppose  $X$  is a random variable with uniform distribution on the interval  $[0,1]$ , often denoted  $X \sim Uniform[0,1]$ . Then

$$f(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

Suppose we are now asked to calculate the probability that  $X$  falls between .3 and .7, i.e.,  $P(.3 < X < .7)$ . We simply integrate the pdf over the values specified by this event:

$$P(.3 < X < .7) = \int_{.3}^{.7} f(x) dx = \int_{.3}^{.7} 1 dx = .7 - .3 = .4.$$

The mean, variance, and standard deviation are also straightforward to find:

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} x f(x) dx = \int_0^1 x \cdot 1 dx = \frac{1}{2} x^2 \Big|_{x=0}^{x=1} = \frac{1}{2}(1 - 0) = \frac{1}{2} \\ Var(X) &= E[X^2] - (E[X])^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12} \\ \sigma_X &= \sqrt{Var(X)} = \frac{1}{\sqrt{12}}. \end{aligned}$$

### Example 2

Suppose that  $X$  has an exponential distribution with pdf

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Here  $\lambda$  is a parameter value satisfying  $\lambda > 0$  that changes the shape of the probability distribution. Thus there is a whole family of exponential distributions, just like there is a family of binomial and Poisson distributions. Note that  $X$  can only take positive values.

Calculating the probability of events is simple. If  $a > 0$ , then

$$P(a < X < b) = \int_a^b f(x) dx = \int_a^b \lambda e^{-\lambda x} dx = -e^{-\lambda x} \Big|_{x=a}^{x=b} = e^{-\lambda a} - e^{-\lambda b}.$$

The mean and variance are theoretically simple, but a bit more complicated to do by hand, requiring integration by parts (or use of a good table or mathematical software package like Maple or Mathematica). It is sufficient for you to know these results and that you theoretically could compute the following:

$$\begin{aligned} E[X] &= \frac{1}{\lambda} \\ \text{Var}(X) &= \frac{1}{\lambda^2}. \\ \sigma_X &= \frac{1}{\lambda}. \end{aligned}$$

Thus the mean is equal to the standard deviation (not to be confused with the Poisson, where the mean is equal to the variance).

## 4 Cumulative Distribution Functions

For any random variable  $X$ , the *cumulative distribution function*, or CDF, is defined as  $F(x) = P(X \leq x)$ , which is a function of the cut-off value  $x$ . For a discrete random variable, this is a sum,  $F(x) = \sum_{x_i < x} p_i$ , and for a continuous random variable, this is an integral,  $F(x) = \int_{-\infty}^x f(t) dt$ . You should recognize this integral as just finding the probability of the infinite interval from  $-\infty$  to  $x$ , the integral of the pdf over that interval. Note that we use a different letter, like  $t$ , as the indexing letter for the pdf, since we then integrate out  $t$  and evaluate the result at  $x$ .

### Example 3

Suppose  $X$  is a discrete random variable with a probability distribution given by the following table:

$x$	0	1	2
$P(X = x)$	0.2	0.3	0.5

Then its CDF is

$$F(x) = P(X \leq x) = \begin{cases} 0 & \text{if } x < 0 \\ 0.2 & \text{if } 0 \leq x < 1 \\ 0.5 & \text{if } 1 \leq x < 2 \\ 1 & \text{if } x \geq 2. \end{cases}$$

Note that the CDF is defined as a function over the entire real line, so we need to specify that it is zero for values below the valid range of the possible values, that once it starts hitting possible values, it stays constant in between values, and that it is one above the range of possible values.

### Example 4

Suppose  $X$  is uniformly distributed on the interval  $(a, b)$ , where  $a$  and  $b$  are real numbers and  $a < b$ . We denote this  $X \sim \text{Unif}(a, b)$ . The pdf for  $X$  is  $f(x) = \frac{1}{b-a}$  if  $a < x < b$  and zero otherwise (notice that this integrates to 1). Thus the CDF is

$$F(x) = \int_{-\infty}^x f(t) dt = \begin{cases} 0 & \text{if } x \leq a \\ \int_a^x \frac{1}{b-a} dt = \frac{x}{b-a} - \frac{a}{b-a} = \frac{x-a}{b-a} & \text{if } a < x < b \\ 1 & \text{if } x \geq b. \end{cases}$$

Again, note that we must specify that it is zero below  $a$  and one above  $b$ .

## 5 Exercises

- Suppose  $X$  is a random variable with pdf  $f(x) = \begin{cases} cx & \text{if } 0 \leq x \leq 4 \\ 0 & \text{otherwise.} \end{cases}$ 
  - Find  $c$  (recall that the total probability is always 1). Use this value of  $c$  for the rest of the calculations in this problem.
  - Find  $P(-1 \leq X \leq 1)$ .
  - Find  $P(X > 2)$ .
  - Find  $P(X < 3 | X > 1)$ .
  - Find  $E[X]$ .
  - Find  $\text{Var}(X)$  and  $\sigma_X$ .
  - Find the CDF of  $X$ ,  $F(x)$ .
- Suppose  $X$  is a random variable with pdf  $f(x) = \begin{cases} \frac{3}{2}x^2 & \text{if } -1 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$ 
  - Find  $P(-2 < X < 0)$ .
  - Find  $P(X > -.5)$ .
  - Find  $P(X > .5 | X > -.5)$ .
  - Find  $E[X]$ .
  - Find  $\text{Var}(X)$  and  $\sigma_X$ .
  - Find the CDF  $F(x)$ .
- Suppose  $X$  is a random variable with pdf  $f(x) = \begin{cases} x & \text{if } 0 < x < 1 \\ 2 - x & \text{if } 1 \leq x < 2 \\ 0 & \text{otherwise.} \end{cases}$ 
  - Find  $P(1 < X < 3)$ .
  - Find  $P(X > .5)$ .
  - Find  $P(X > 1 | X > .5)$ .
  - Find  $E[X]$ .
  - Find  $\text{Var}(X)$  and  $\sigma_X$ .
  - Find the CDF  $F(x)$ .
- Suppose  $X$  is an exponential random variable with mean 4.
  - Find the pdf of  $X$ .
  - Find  $P(1 < X < 3)$ .
  - Find  $P(X > 3)$ .
  - Find  $P(X > 6 | X > 3)$ .
  - Find  $\text{Var}(X)$  and  $\sigma_X$  (use the formula; don't do the integral).
  - Find the CDF  $F(x)$ .